Internet Digital Library System

Author: Hyungjoo Em
email: hem@mcs.kent.edu, homepage: http://www.mcs.kent.edu/~hem
Prepared for Prof. Javed I. Khan
Department of Math and Computer Science, Kent State University
Date: *May 1999*

Abstract: Digital Library System(DLS) is a system that manages a very large volume of various types of digitized data. Because DLS uses computers to build, to manage, and to search the contents, it requires efficient search tools. Recent interest in DLS originated from the vision of Digital Library Initiative, which was jointly sponsored by NSF, ARPA, and NASA in 1994(http://iita.ivv.nasa.gov/program/iita1997/dlt/dlt_jri.html).

This survey covers the issues of Internet Digital Library System and Solutions according to its data type and contents. This survey also introduces some examples of the typical Digital Library Systems.

Other Survey's on Internetwork-based Applications
Back to Javed I. Khan's Home Page

Table of Contents:

Basic Mechanism

Digital Library refers to the library implemented using the computer systems connected through the network(s). All the information in the DLS is available on-line to the client.

As we see in Figure1, when

1. clients request the data through the Internet,
2. the web server will route the request to the *search and management module* (or process), and
3. searching is performed on the meta-database (meta-index) to generate result for clients. Then according to the client response
4. actual data will be retrieved from media database.

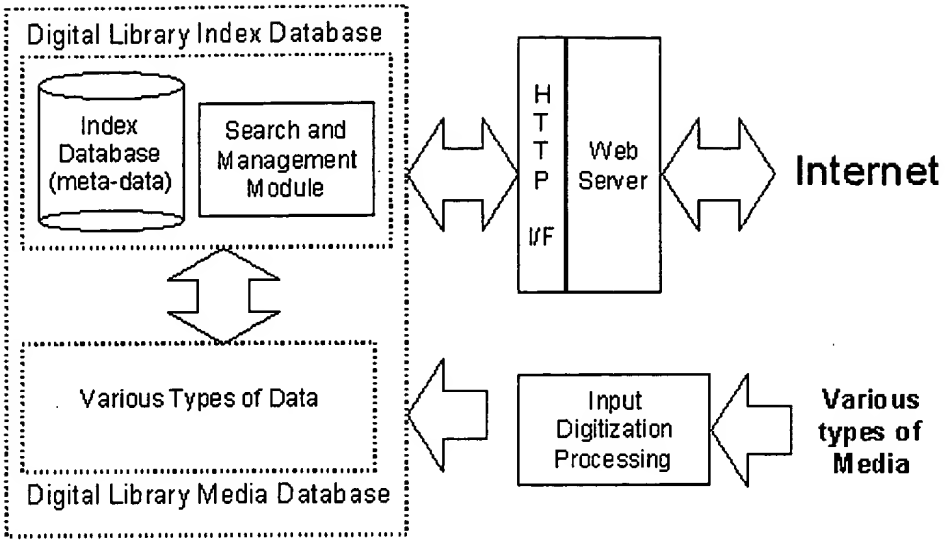

Figure 1 Internet Digital Library System(DLS) Architecture

Problems and Approach

TWO MAJOR CHARACTERISTICS.

Although DLS can be viewed in many different ways, in general, it can be defined as a very large collection of digitized data measured by the unit of Terabytes. DLS also has a variety of data types: text as a basic data type, image, audio, and video.

1. The way to manage large amount of digitized multimedia data effectively and how to make it available for users. This will involve huge amount of works for digitization process.
2. Requirement of efficient and various search tools because all types of data from various fields can be integrated into DLS.

ISSUES AND SOLUTIONS

Digitization of Library Text Data (Books, Journals, Papers, and so forth)

Most of currently available DLSs are physical library data that are translated into digital form routinely. For example, books, journals, and movies are all examples of physical library data that are scanned, digitized, or otherwise translated and make it available through the computer network(on-line).

A major problem is to decide when translating a book into digital form,

1. When does an ASCII representation of the text suffice?
2. When must each page be scanned as a photograph would be?

According to their use conversion process varies. We have to take the characteristics of physical data into consideration, for example, size and shape of the book, pictures in the book. We have to preserve the characteristics of the book as closely as possible.

Quality Loss by Digitization Process of Multimedia Data :

Since it is *digital* we have to convert existing physical data into digital form.

When we are converting multimedia data - images, video, and audio - into digital form, we are introducing data loss in a variety of forms. That data loss includes:

- Loss by digitization
- Loss by compression

DLS Example – Reflection of Cultures

Different culture has different requirements.[1] Brief look at the Digital Libraries currently available which reflect different cultures. These are all Imaging Libraries.

- *The Vatican Library*: because of the limited environment of the library, team is organized to provide online digital library services to scholars. This involves scanning(digitization of printed/hand written materials) which requires high quality of images for scholarly research
    - 2600 x 300 pixels for each picture, compressed by JPEG.
    - Psychological concern about substitutions for original images by digitized data
    - For the security reason, put technology visible watermark.
- *The (U.S.) National Gallery of Art*: Requirement was achieving the highest practical image quality (http://www.nga.gov).
    - Scanning from transparency to 3000 x 4000 pixels resolution.
    - Retouching was need to make up for the color differences and noises.
    - Regenerate images for public(www) use into 610x390 pixels → SUCCESS!
- *The Klau Library of the Hebrew Union College*: Nearly 750,000 volumes, one of the largest collections of Judaica. Its Dalsheimer Rare Book Exhibit includes rare incunabula, illuminated manuscripts, Biblical codices, communal records, legal documents, and scientific tracts.
    - Image capture and preparation technologies.
    - Watermarks for access on the Internet.
- The State Hermitage Museum: Comprehensive set of museum information services including, Web site, an Education center, and a network of information kiosks (http://www.ibm.com/stories/1997/06/art3.htm).
    - Direct image capture of small three-dimensional objects, retouching of images captured from photographic materials.

DLS – Usage as the Substitute for Originals

Where the captured Images are used as a substitution for originals, the images should meet the needs required for its use while keeping the size and quality loss minimum. Here is the way to produce high-quality images for Internet Distribution[5].

- *Accurate Color*: Color reproduction is highly dependent upon the equipment for example, scanner for capturing, monitor for displaying, and printer for printing.
    - Standard Color Profile(http://www.color.org/) for device-independent color.
    - Strategy 1: Color-metric utilizes colored filters.
    - Strategy 2: Color filters multidimensional interpolation based on the device color profile(the

most popular)
- Strategy 3: Displaying the image on a monitor, involves retouching the image, and matching the monitor's color profile with the image.

- *Sharp Images*: Images obtained by capturing devices are generally blurred and may have some aliasing artifacts.
  - Capture images in a higher resolution and reduce it later.
  - Sharpening increases the size of compression result, it greatly improve perception.

- *Noises*: Noise will lead to larger compressed data sizes. It is preferable to use longer capture time to produce higher quality and higher compression rate instead of having fast scan and lower compression rate.

Searching in the Multimedia DLS

Prevailing images in the Internet today demand the efficient image searching as text searching. In a typical environment of content-based search, search system uses descriptors to search images.

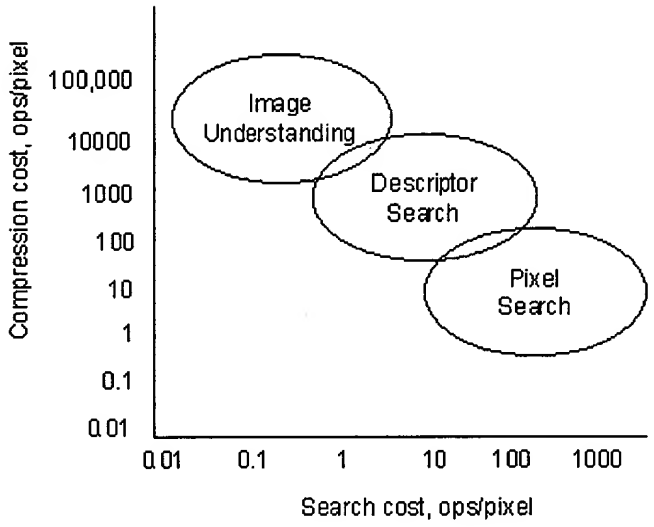

Figure 2 Trade-offs between compression cost and search complexity[6]

Figure 2 shows how the Search cost and Compression cost varies according to the searching level(scale is logarithmic).

- Pixel level search: costly
- Image understanding: we need compression and decompression cost for visual interpretation.
- Descriptor Search: we can make it more efficient by building meta-database for images.
  For the same idea, for the Digital Library for Images by building and using descripter database, we can perform efficient Digital Library Search.

Metadata Creation for Efficient Searching

We need different way for searching multimedia data. Multimedia data itself does not have any useful information unless we retrieve the data for visualization. Therefore, we can use external database to store the meta-data for searching.>

Searching in Text-Based DLS

Searching data in a DLS is the most important issue.

Main issue for text data type falls into efficient searching algorithm. Because internet search returns too much information, *reliable ranking algorithm* of result is required. Some of the researches for ranking algorithms are based on followings.

- Using the text for links to other pages as descriptors of the page.
- Based on the number of previous users' click from the result(www.directhit.com).
- Based on the number of links TO good pages(hubs), and number of good links FROM good link pages(authorities).

Another approach to efficient searching is called *Meta-search* A metasearch engine searches the Web by making requests to multiple search engines. This can combine the result of multiple search engines and the is able to provide a consistent user interface for searching these engines.

If we have algorithm to combine the search result performed by various search engines we can get the best search result.

- NEC Research Institute's Inquirus : Download and analyze the individual documents. Improve by providing more useful document summaries.
- Query Sensitive Search: Providing query sensitive summaries allows users to judge accurately and rapidly.
- Use different Summarization Strategy for different types of data.

Unsolved Problems and Difficulties.

NETWORK BANDWIDTH

Contrast to the earlier pace of Internet, now we have hundreds of millions of global network nodes, which means billions of people are accessing Internet everyday.

Current Internet reaches its limit. Even though the Internet caching technologies have been developed, they are just to avoid the Internet traffic. They do not solve the fundamental bandwidth problem. So it is preferable to design a new generation global network of high bandwidth. Here are the current projects of new Internet.

- Next Generation Internet(NGI) (http://www.ngi.gov)
- Internet 2 (http://internet2.edu and http://www.ucaid.edu)

Backbone bandwidth will be 10Gb/2 (OC192) or faster, hundreds of GigaPOPs, and Quality-of-Service mechanism. DLS requires fundamental changes to the Internet network infrastructure and in the near future DLS will be the prime application.

DIGITIZATION

We have to face a certain amount of quality loss to implement Digital Library Systems on current Internet infrastructure. What quality is good enough for a Digital Library System and how to measure the quality.

If we perform a descriptor search, what strategy should we use to construct descriptor database(meta-

data). And also what mechanism should we use to update and replace descriptor data.

How to get Exact Color Information and what are the criteria?

MAINTENANCE

Maintenance will be critical to digital libraries, especially promoting diverse and broad informal materials. When ignored, maintenance issues within the digital library will threaten its usefulness and even its long-term

Maintenance problem can be categorized as being both technical and institutional,

library as both institution and technology. We have to examine software architecture , type of contents, and dynamic contents that contain informal and dynamic material will have substantially greater maintenance problems.

KEEP DATA TRANSPARENCY TO APPLICATION

The underlying architecture should be separate from the content stored in the library.

A conventional research library stores more than books and the digital library are the same. Almost every type of information can be represented in digital form, including text, pictures, musical works, computer programs, databases, models and designs, video programs, and compound works combining many types of information.

This underlying architecture is a base for extensions that can be tailored for various types of information. The extensions typically include specific formats, protocols, and rights management that are appropriate for the type of material. For example, the extensions for digitized movies will be very different from those for video games.

Digitization cost is extremely high and once the data is digitized it can be used over and over again. However the application need can be changed. So if we do not have our digitized data separated from the applications in worst case we have to re-digitize the whole data.

More to think about.

INTERACTIVE ENVIRONMENT[9]

Conventional libraries are interactive environments. It is a meeting places for communication and collaboration. The use of conventional libraries might be highly interactive and collaborative.

These observations call for corresponding tools and support in virtual libraries. Communication in virtual libraries, it introduces well known forms of communication and shows whether and how these are found in virtual libraries.

INTEGRATING DIFFERENT PERSPECTIVES ON DESIGN RATIONALE[8]

Digital Libraries System is designed by Design rationales of the people.

To some it implies argumentation and frameworks for argumentation. To others it implies the

documentation of design, like that required for many types of industrial or government work. There should be an integrated approach to design rationale where design communication is captured and, over time, incrementally structured into argumentation and other formalisms to enable improved retrieval and use of this information.

Reference and Pointer to more information.

References

1. Fred Mintzer, IBM Thomas J. Watson Research Center, Developing Digital Libraries of Cultural Content for Internet Access, IEEE Communications Magazine, January1998,pp. 72-78.
2. S. Lawrence and C.L. Giles, *Searching the Web: General and Scientific Information Access*," IEEE Communications Magazine, January1998, pp. 116-122.
3. S. Lawrence and C.L. Giles, "Context and page analysis for improved Web search," IEEE Internet Comp., vol. 2, no.4, 1998, pp. 38-46.
4. Peter J. Nürnberg, Richard Furuta, John J. Leggett, Catherine C. Marshall, and Frank M. Shipman III, "Digital Libraries: Issues and Architectures", 1995. (http://www.csdl.tamu.edu/csdl/DL95/papers/nuernberg/nuernberg.html)
5. B. A. Wandel, *Foundations of Vision*, Sunderland, MA: Sinaur Associates, 1995.
6. Harold S. Stone, NEC Research Institute, *Image Libraries and the Internet*, " IEEE Communications Magazine, January1998, pp. 116-122.
7. *Next Generation Internet(NGI)*, (http://www.ngi.gov)
8. Frank M. Shipman III, Raymond J. McCall, *Integrating Different Perspectives on Design Rationale: Supporting the Emergence of Design Rationale from Design Communication*, (http://www.csdl.tamu.edu/~shipman/aiedam/)
9. Klaus Tochtermann, A First Step Toward Communication in Virtual Libraries, (http://www.csdl.tamu.edu/csdl/pubs/klaus/TecRepKlaus.html)
10. International Journal of Digital Library (http://www.informatik.uni-trier.de/~ley/db/journals/jodl/)
11. D-*lib* Magazine (http://www.dlib.org/)
    D-Lib Magazine is a monthly magazine about digital libraries for researchers, developers, and the intellectually curious. New issues are published on the 15th of each month.

Research Institutes

- *The Center for the Study of Digital Libraries (CSDL) – Texas A&M University (http://www.csdl.tamu.edu/)*
  CSDL was established in 1995 by The Texas A&M University System Board of Regents. The web site contains publications and conference materials.

- *NSF-DARPA-NASA Digital Library Joint Research Initiative (http://iita.ivv.nasa.gov/program/iita1997/dlt/dlt_jri.html)*

  This project includes 1) UC Santa Barbara, UC Berkeley, Stanford University for interoperability, and 2) Carnegie Mellon University, University of Illinois, and Unviersity of Michigan for searching.

Example Sites

- *Microsoft Terra Server (Satellite Image Database)*
  (http://www.terraserver.com/→ Front page of TerraServer)
  (http://terraserver.microsoft.com/default.asp → TerraServer Startup Page)

Microsoft TerraServer manages satellite images from all over the world and lets you [redacted] buy high-resolution satellite images.

- *ACM Digital Library*
  (http://www.acm.org/dl/)

  Digital Library managed by ACM contains bibliographic information, citations, and fulltext articles.

- ** Also see the implementation examples from Schools of NSF-DARPA-NASA Digital Library Joint Research Initiative.*

Products

- *IBM DB2 Digital Library Version2*
  (http://www.software.ibm.com/is/dig-lib/)